

SECTION III: DEVELOPMENT AND REPORTING OF SCORES

CHAPTER 11—SCORING

MACHINE SCORED ITEMS

Once the 1998–99 booklets had been logged in, identified with appropriate scannable, pre-printed school information sheets, examined for extraneous materials, and batched, they were moved into the scanning area. For all response booklets (and questionnaires and other forms that require imaging/scanning) to be imaged, this area is the last stop in the processing loop in which the documents themselves are handled.

At that point, 100% of the response document and other scannable information necessary to produce the required reports had been captured and converted into an electronic format, including all student identification and demographics, selected-response answers, and digital image clips of hand-written responses. The digital image clip information allowed Advanced Systems to replicate student responses just as they appeared on the originals, but they had been transferred onto the readers’ monitors. From that point on, the entire process—data processing, scoring, “range-finding,” data analysis, reporting—was accomplished without further reference to the originals.

The first step in that conversion was the removal of the booklet bindings so that the individual pages could pass through the scanners, one at a time. Once cut, the sheets were put back in their proper boxes and placed in storage until needed for the scanning/imaging process.

Customized scanning programs for all scannables were prepared to selectively read the student response booklets and to format the scanned information electronically according to pre-determined requirements. Any information (including multiple-choice response data) that had been designated time-critical or process-critical was handled first.

In addition to numerous real-time quality control checks, duplex read, a transport printer that prints a unique identifying number on each sheet of each booklet, and on-line editing capability, the new 5000i scanners offer features that make them compatible with Internet technology.

SCANNING QUALITY CONTROL

NCS scanners are equipped with many built-in safeguards that prevent data errors. The scanning hardware is continually monitored for conditions that will cause the machine to shut down if standards are not met. It will display an error message and prevent further scanning until the condition is corrected. The areas monitored include document page and integrity checks, user-designed on-line edits, and many internal checks of electronic functions.

Before every scanning shift begins, Advanced Systems' operators performed a daily diagnostic routine. This is yet another step to protect data integrity, and one that has been done faithfully for the many years that we have been involved in production scanning. In the rare event that the routine detects a photocell that appears to be out of range, we calibrate that machine and perform the test again. If the read is still not up to standard, we call for assistance from our field service engineer.

As a final safeguard, spot checks of scanned files, bubble by bubble and image by image, were routinely made throughout scanning runs. The result of these precautions, from the original layout of the scanning form to the daily vigilance of our operators, was a scan error rate well below 0.001.

ELECTRONIC DATA FILES

Once the data had been entered and the scanning logs and other paperwork completed, the booklets themselves were put into storage (where they stayed for at least 180 days beyond the close of the fiscal year). When it had been determined that the files were complete and accurate, those files were duplicated electronically and made available for many other processing options. Completed files were loaded onto our local area network (LAN) for transfer to

Advanced Systems' proprietary I-Score system for scoring. Those files were then used to identify (and print out) papers to be used in the rangefinding and standard-setting processes, and the data was made transferable via the Internet, CD-ROM, or optical disk.

ITEMS SCORED BY READERS

Test and answer materials were handled as little as possible to minimize the possibility of loss, mishandling, or breach of security. Once scanned, either by optical mark reader or the I-Score system, papers were stored securely in areas with limited personnel access.

As explained in the following sections on scoring, the I-Score system itself ensures the security of responses and test items: all scoring is "blind"; that is, no student names are associated with viewed responses or raw scores and all scoring personnel are subject to the same nondisclosure requirements and supervision as regular Advanced Systems staff.

I-SCORE

After the 1998–99 test material had been loaded into the LAN, I-Score sent electronically scanned images of student work to individual readers at computer terminals who evaluated each response and recorded each student's score via keypad or mouse entry. When the reader had finished with one response, the next response appeared immediately on the computer screen. In that way, the system guaranteed complete anonymity of individual students and ensured the randomization of responses during scoring.

Although I-Score is based on conventional scoring techniques, it also offers numerous benefits, not the least of which is raising the bar on scoring process capability. Some of the benefits are as follows:

- real-time information on scorer reliability, read-behinds, and overall process monitoring;
 - early access to subsets of data for tasks such as standard setting;
 - reduced material handling, which not only saves time and labor, but also enhances the security of materials;
- and

- immediate access to samples of student responses and scores for reporting and analysis through electronic media.

Scoring operations, directed by the manager of scoring services, are carried out by a highly qualified staff. The staff included:

- chief readers, who oversaw all training and scoring within particular subject areas;
- quality assurance coordinators (QACs), who lead rangefinding and training activities and monitor scoring consistency and rates;
- verifiers, who perform read-behinds of readers and assist at scoring tables as necessary; and
- readers, who perform the bulk of the scoring.

Table 11-1 summarizes the qualifications of the 1998/99 MEA quality assurance coordinators and readers.

| Table 11-1 Qualifications of 1998/99 MEA QACs and Readers | | | | | |
|--|-------------------------|---------|-----------|-------|-------|
| Scoring Responsibility | Educational Credentials | | | | Total |
| | Doctorate | Masters | Bachelors | Other | |
| QACs | 11% | 68% | 21% | --- | 100% |
| Readers | 2% | 17% | 35% | 46% | 100% |

PRELIMINARY ACTIVITIES

Preliminary activities for scoring included (1) participating in the planning and design of documents to be used for scoring, (2) reviewing items and score guides for rangefinding and training and the creation of rangefinding packets, and (3) selecting scoring staff and training them for scoring.

PLANNING AND DESIGNING DOCUMENTS

At the request of Advanced Systems' project manager, scoring personnel advised project management and DOE staff on the program design in order to support an efficient and effective scoring process. Scoring staff contributed also to the design of

- response documents and the image-capture process to yield acceptable image clips (also defining file format and layout); and
- scoring benchmarks composed of the guide, subject background information, and anchor papers.

REVIEWING ITEMS AND GUIDES (RANGEFINDING)

Before the scheduled start of scoring activities, scoring center staff reviewed test items and scoring guides for rangefinding. At that point, chief readers and selected QACs prepared scorer training materials.

Advanced Systems' scoring staff (including test developers) selected one or two anchor examples for each item score point. An additional six to ten responses per item were chosen as part of the training pack. The anchor pack consisted of mid-range exemplars, while the training pack exemplars illustrated the range within each score point. The chief readers, who worked closely with QACs for each content area, facilitated the selection of response exemplars. One of the greatest difficulties in the selection of anchor and training exemplars was finding a sufficient number of papers representing the highest scores (4 or 8) as such scores are fairly rare.

SELECTING AND TRAINING SCORING STAFF

SELECTING QUALITY ASSURANCE COORDINATORS (QACs) AND VERIFIERS

Because the read-behinds performed by the QACs and verifiers moderated the scoring process and thus maintained the integrity of the scores, individuals to fill those positions were selected for their accuracy. In addition, QACs, who train readers to score each item in their content areas, were selected for their ability to instruct and for their level of expertise in their content areas. For this reason, QACs typically are retired teachers who have demonstrated a high level of expertise in their respective disciplines. The ratio of QACs and verifiers to readers was approximately 1:11.

TRAINING QUALITY ASSURANCE COORDINATORS AND VERIFIERS

To ensure that all QACs provided consistent training and feedback, the chief readers spent two days training and qualifying the QACs, and the QACs reviewed all items with the verifiers before scoring. In addition, QACs rotated

among tables, supervising readers and reading behind verifiers, who in turn read behind a different table of readers each day.

SELECTING READERS

Applicants were required to demonstrate their ability by participating in a preliminary scoring evaluation. The I-Score system enables Advanced Systems to efficiently measure a prospective reader's ability to score student responses accurately. After having participated in a training session, applicants were required to achieve at least 80% exact scoring agreement for a qualifying pack consisting of 20 responses to a predetermined item in their content area. Those 20 responses were randomly selected from a bank of approximately 150, all of which had been selected by QACs and approved by the chief readers and developers.

TRAINING READERS

The QACs first applied the language of the scoring guide for an item to its anchor pack exemplars. Once discussion of the anchor pack had concluded, readers attempted to score the training pack exemplars correctly. The QACs then reviewed the training pack and answered any questions readers had before actual scoring began. With this system, two aspects of scoring efficiency are in conflict. First, in order to minimize training expense, it is desirable to train each reader on as few items as possible. Second, to prevent reader drift and to minimize retraining requirements, it is desirable to score a given item in a brief period of time. But the lower the number of unique items each reader scores, the greater the number of readers required to score that item quickly. To minimize that conflict, we divided each subject area's readers into two or more groups. On the first day of scoring, each group was trained to score a different item. When a group had completed all of an item's responses, those readers were trained on another item (or set).

SCORING ACTIVITIES

Student test booklets at grade level 4 and student response booklets at grade levels 8 and 11 were digitally scanned and scored on a file server for a dedicated, secure LAN. I-Score then distributed digital images of student responses to readers. Training and scoring took place over a period of approximately two weeks.

Items were randomly assigned to readers; thus, each item in a student's response booklet was more than likely scored by a different reader. By using the maximum possible number of readers for each student, the procedure effectively minimized error variance due to reader sampling. Matrix writing prompts, as well as all common and matrix-constructed and extended-response items were scored once with a 2% read-behind to ensure consistency among readers and accuracy of individual readers.

MONITORING READERS

After a reader scored a student response, I-Score determined whether that response should also be scored by another reader, scored by a QAC or verifier, or routed for special attention. QACs and verifiers used I-Score to produce daily reader accuracy and speed reports. QACs and verifiers were able to obtain current reader accuracy reports and speed reports on-line at any time.

Two readers scored some responses to open-response items. The weighted average of exact (both readers assigned the paper the same score) and adjacent (the two readers scores differed by one point) percent agreement of double-blind scores are reported below. Averages were weighted by the number of papers that were double-blind scored for each item. Additionally, the weighted average of the correlation of scores provided by the two readers is reported for each subject and grade combination. Correlations were Fisher z transformed, weighted by the number of papers double-blind scored for each item, then averaged. The average z was transformed back to the correlation metric. Up to approximately 20% of the responses for each item received double-blind scores.

| Table 11-2 Interrater Consistency of Common Open Response Items | | | |
|--|----------------|--|--|
| Grade | Subject | Average Percentage of Exact and Adjacent Agreement | Average Correlation of First and Second Scores |
| 4 | Reading | .99 | .75 |
| | Mathematics | .97 | .89 |
| | Science | .99 | .81 |
| | Social Studies | .98 | .73 |
| 8 | Reading | .99 | .71 |
| | Mathematics | .99 | .95 |
| | Science | .99 | .83 |
| | Social Studies | .99 | .79 |
| 11 | Reading | .99 | .67 |
| | Mathematics | .95 | .89 |
| | Science | .97 | .84 |
| | Social Studies | .96 | .79 |

SCORING THE WRITING

Maine teachers and administrators were recruited to score the common writing prompt at in-state scoring sessions that were held in Bangor and Gorham, Maine. Teachers who participated in the scoring process developed skills in holistic evaluation of writing using a rubric aligned with the standards outlined in the Maine *Learning Results*. Those skills could then be applied to writing instruction in the classrooms, and the scoring of writing also gave participants an opportunity to read the range of student writing produced at each grade and to connect their current teaching practices with the recommendations in the Maine *Learning Results*. Administrators who participated gained skills helpful in improving the teaching and evaluation of writing in their schools. Maine teachers' involvement in scoring also created a network of teachers who served as a resource to their local and state schools.

GENERAL SCORING GUIDES

SHORT-ANSWER ITEMS

| Score Point | Description |
|-------------|---|
| 2 | ▪ The student's response provides a complete and correct answer. |
| 1 | ▪ The student's response is partially correct. ▪ The student's response may be incomplete or contain errors. |
| 0 | ▪ The student's response is totally incorrect or too minimal to evaluate. |
| B | ▪ Blank/no response. |

OPEN RESPONSE ITEMS

| Score Point | Description |
|-------------|---|
| 4 | ▪ The student completes all important components of the task and communicates ideas clearly. ▪ The student demonstrates in-depth understanding of the relevant concepts and/or processes. ▪ When instructed to do so, the student chooses more efficient and/or sophisticated processes. ▪ When instructed to do so, the student offers insightful interpretations or extensions (e.g., generalizations, applications, and analogies). |
| 3 | ▪ The student completes the most important components of the task and communicates clearly. ▪ The student demonstrates understanding of major concepts even though he/she overlooks or misunderstands some less important ideas or details. |
| 2 | ▪ The student completes most important components of the task and communicates those clearly. ▪ The student demonstrates that there are gaps in his/her conceptual understanding. |
| 1 | ▪ The student shows minimal understanding. ▪ The student addresses only a small portion of the required task(s). |
| 0 | ▪ The student's response is totally incorrect or irrelevant. |
| B | ▪ Blank/no response. |

WRITING PROMPTS

Stylistic & Rhetorical Aspects of Writing Topic Idea Development

| 1 | 2 | 3 | 4 | 5 | 6 |
|--|---|---|--|---|---|
| <ul style="list-style-type: none"> Little topic development and/or organization, few details Possible evidence of voice Simplistic language (wording and sentence structures) | <ul style="list-style-type: none"> Limited topic development, focus, and/or details Evidence of voice Limited variety in language used (wording and sentence structures) | <ul style="list-style-type: none"> Moderate topic development, focus, and details Some voice Some variety in language used (wording and sentence structures) | <ul style="list-style-type: none"> Well developed with control and relevant details Consistent voice Variety in language used (wording and sentence structures) | <ul style="list-style-type: none"> Fully developed with strong details Sustained voice and/or tone with emerging style Effective use of language | <ul style="list-style-type: none"> Topic and details richly developed Distinctive voice, tone and style Rich use of language |

Analytic Annotations

| | | Commendations | Needs |
|--------------------------|--|--|---|
| Topic Development | The overall effect of the paper | TX sustained development throughout TY creative, insightful, and/or shows voice | TJ less repetition of ideas TK more development of ideas/topic |
| Organization | The degree to which the response is: <ul style="list-style-type: none"> Focused Clearly and logically ordered Clarified by paragraphs | OX clearly focused OY logical organization of ideas | OJ clearer focus OK more effective use of paragraphing |
| Details | The degree to which the response includes examples that develop the main points | DX details support focus DY uses interesting details | DJ to avoid simply listing details DK more/relevant details |
| Language/Style | The degree to which manipulation of language, including vocabulary, word choice, word combination, and sentence variety is effective | LX word choice enhances meaning LY sentence variety is used effectively | LJ more attention to word choice LK more variety in language |

Standard English Conventions

| 1 | 2 | 3 | 4 |
|---|--|--|--|
| <ul style="list-style-type: none"> Errors seriously interfere with communication and/or Little control of sentence structure, grammar and usage, and mechanics in first writing draft | <ul style="list-style-type: none"> Errors interfere somewhat with communication and/or Few or no errors in simplistic or limited text in first writing draft | <ul style="list-style-type: none"> Errors do not interfere with communication and/or Few errors relative to length of essay or complexity of sentence structure, grammar and usage, and mechanics in first draft writing | <ul style="list-style-type: none"> Control of a variety of sentence structures, grammar and usage, and mechanics Length and complexity of essay provide opportunity for student to show control of standard English conventions in first draft writing |

Analytic Annotations

| | | Commendations | Needs |
|--------------------------|---|--|---|
| Sentences | The degree to which the response includes sentences that are correct in structure | SP correct sentence structure | SR correct sentence structure |
| Grammar and Usage | The degree to which the response demonstrates correct <ul style="list-style-type: none"> Use of standard grammatical rules of English Word usage and vocabulary | GUP correct application of grammatical rules GUQ control of vocabulary and word usage | GUR correct application of grammatical rules GUS greater attention to correct word usage |
| Mechanics | The degree to which the response demonstrates correct <ul style="list-style-type: none"> Punctuation Capitalization Spelling | MP control of mechanics aids clarity MQ correct mechanics in sophisticated construction | MR greater control of mechanics MS more careful editing |

